Södertörns högskola | Institutionen för livsvetenskaper
Magisteruppsats 30 hp | Bioinformatik | andra terminen 2009

Adaptive evolution of Transcription Factors in European and wine yeast

Av: John Boss
Handledare: Anthony Wright och Mats Grahn

södertörns högskola
SÖDERTÖRN UNIVERSITY

Abstract

The mutability of transcription factors (**TF**) is thought to be of high importance for the evolutionary change of living organisms. Transcription factors, coactivators, coregulators, kinases, chromatin remodelers conditional factors and other proteins together govern the timing and level of gene expression. About 10% of the genes in the human genome are predicted to be TFs [1] and mutational changes in these genes or in the target regulatory sequences they bind will potentially give rise to evolutionary advantages or malfunctions for the organism. Recent research has suggested that the parts of the transcription factors that are not structurally defined in solution, so called intrinsically disordered regions (IDRs), have a higher potential for evolutionary diversification than more structurally rigid regions. This suggests that these domains that earlier have been considered mostly unimportant may have an important potential for evolutionary diversification. This project aimed to further evaluate evidence supporting the hypothesis that variable-structured domains in transcription factors are of significant importance for functional diversification. This was be done by comparing the rate of synonymous and non-synonymous genetic variation in the coding regions of 12 selected TFs within a highly conserved clade of European wine yeasts and by comparing this variation to divergent phenotypic patterns within the strains. The frequency of non-synonymous mutations was much greater than for synonymous mutations indicating an important role of positive selection acting on these TFs during diversification of the different strains. No significant connections were discovered between the distribution of DNA variation and phenotypic patterns.

(Keywords: Transcription factors, intrinsically disordered regions, yeast, wine, positive selection)

Introduction

Intrinsically disordered regions in complex organisms

Intrinsically disordered regions (IDRs) are domains in proteins that do not have a native stable structure in isolation. They need to bind to other proteins or macromolecules for any measurable structure to appear. A protein can consist of any percentage of IDR, from fully unstructured, partly folded to totally stable [2]. It has been shown that proteins that are rich in IDRs are most often found in the cellular nucleus, where they are involved in signalling systems and transcriptional regulation. Generally it has been shown that more complex organisms have, in proportion, more IDR-domains than simpler organisms. When predicting IDR regions using bioinformatics (in silco) assays it has been suggested that 30%, 33% or as much as half of the eurkaryotic proteins include IDRs. The value depends on what threshold value is used as minimal sequence length for a IDR domain [2] [3] [4]. In Archea the percentage were calculated to 2.0 and to 4.2% for eubacterial proteins [3].

A Short history of structural biology

In the 20th century the paradigm that protein 3D structure is a prerequisite for its function was widely accepted. In 1894 Fischer presented the theory that yeast enzymes react with glucoside by fitting to each other like a key in a lock. He saw that yeast could hydrolyze α-glucosides but not β- glucosides [5]. Later, in 1936 Mirsk and Pauling showed that the loss of pepsin activity correlated to its denaturation. Edsall speculated in 1952 that maybe all protein functions and interactions could be explained by a relatively small number of fundamental domain structures that are utilised in different combinations in different proteins [2]. IDRs were regarded to be more or less unspecific and neutral in the perspective of evolution because of their unspecific and unstable structure. This improbable conclusion was the result of a wide spread aspiration in the science community to frame the world in a simple, easily understood and clear perspective. This example can be compared to Einstein's distaste for quantum mechanics, and a reminder that we should not impose our simple and beautiful equations on a complicated world, or as Alfred North Whitehead said *"seek simplicity but distrust it*" [6]. However, there existed experimental support for more flexible mechanism of protein function. In 1950 Karush saw examples of

conformational changes in serum albumin and coined the term “configurational adaptability”. Small-angle x-ray scattering further supported that theory in 1978 and 1979 by observation of conformational changes in yeast hexokinase upon binding to glucose [7].

Difficulties when investigating IDR’s

There are many problems involved when investigating IDR-domains, It’s hard to create suitable crystal structures for crystallography when the protein is, by definition, unstable in its pure form. And also the low level of evolutionary conservation at amino acid level makes it hard to find conclusions in comparative studies.
Witin TF’s IDR’s probably function on a more complex and intricate level then stable proteins, bending and binding coactivators and DNA to acquire its functional structure. A high occurrence of mutations may facilitate small changes that regulate how, when and to what extent signalling proteins, RNA Polymerases and DNA interact.

One recent hypothesis is that TFs have a strong potential for evolvability. Evolvable processes provide easy, fast and stable ways to modify cellular processes [8]. Complex organisms, such as ourselves, have a slow generation time and thus also a slow evolutionary adaptability. Minor changes in our living conditions sometimes occur rapidly. It would not be competitive to take many thousands of years to adapt to a more protein rich food intake or create lactose tolerance. To adapt by changing the structure of highly structured enzymes may be a slow mechanism of adaptation. A protein may need more than one mutation to reach a new function and often more than one protein may need to be targeted. Therefore it is natural to consider changes in regulatory systems, such as transcription factors and signalling pathways. A single signalling protein has the ability to change the expression patterns of multiple proteins. In 1990 Michael Conrad declared that biological structure with a high degree of component redundancy and multiple small interactions should have a high degree of adaptability while at the same time remaining functional [9]. An IDR-rich protein could be predicted to function more easily during an intermediate stage of adaptation, where a first mutation may be detrimental while a second mutation may be beneficial. The more dynamic nature of IDR-structures could also provide a bigger spectrum of potential new conformations that are feasible for cellular functions because of a

simple “bend not break design”. One theory is that TFs function as a mixing console/sound board, for the genome, tuning and regulating the different inputs to deliver a balanced and optimal output. In this analogy the IDR would function as the control sliders, giving evolution a flexible way to adjust according to changing conditions [10].

Positive selection in IDR-domains

An earlier study by Johan Nilsson (in prep) investigates the relationship between IDRs and the adaptive evolution of strains belonging to two closely related species of yeast, *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*. Among other things he found an abundance of positive selection in a subset of transcription factors. Positive selection results in an increase in advantageous genetic variation within a population, either in the context of balancing selection, favouring variation, or in the context of a dominating function. And in the level of DNA sequence this can be observed as a higher occurrence of non-synonymous (nucleotides crucial for the amino acid identity in the open reading frame) mutations in relation to synonymous mutations than what would be expected by chance alone (neutral evolution). The opposite is seen in purifying selection where new mutations decrease the fitness of the organism and are therefore selected against by evolutionary pressure.

The aim of this study was to more deeply analyse the evolutionary significance of IDRs. I did this by a focused point study of predicted evolutionary events in 12 transcription factors within a closely related group of 10 European and wine yeast strains. The chosen TFs were YGL254W (also known as FZF1), YLR278C, YML076C (WAR1), YPL230W (USV1), YER184C, YMR280C (CAT8), YJL056C (ZAP1), YMR213W (CEF1), YFL044C (OUT1), YPR104C (FHL1), YBR240C (THI2) and YMR037C (MSN2). These 12 TF were chosen because they had shown high rate of positive selection in the previous study (Johan Nilsson et al, in preparation). The reason for choosing TF’s with already proven high positive selection was to maximise the chance of finding single positive-selected mutations that could account for direct changes in the yeast phenotype.

Several phenotypic tests have been performed in relation to these TFs. For example, YLR278C is located in the nucleus but is not essential and deletion mutants show a distinct growth defect on media containing caffeine [11][12]. YJL056C (ZAP1) is

involved in zinc-response [13] and YML076C (WAR1) induces transcription of PDR12 and FUN34 which encode a putative ammonia transporter and an acid transporter, respectively [14] [15]. YPR104C (FHL1) is involved in rRNA processing [25].

To increase the chance of finding evidence of recent adaptative mutations that can be coupled to functional adaptation, I chose closely related *Saccharomyces cerevisiae* strains derived from a closely related group composed of wine-production related yeast and yeast localised in Europe (Fig1). This ensures that relative few mutations separate the strains apart whilst still subjecting the strains to different environmental influence.

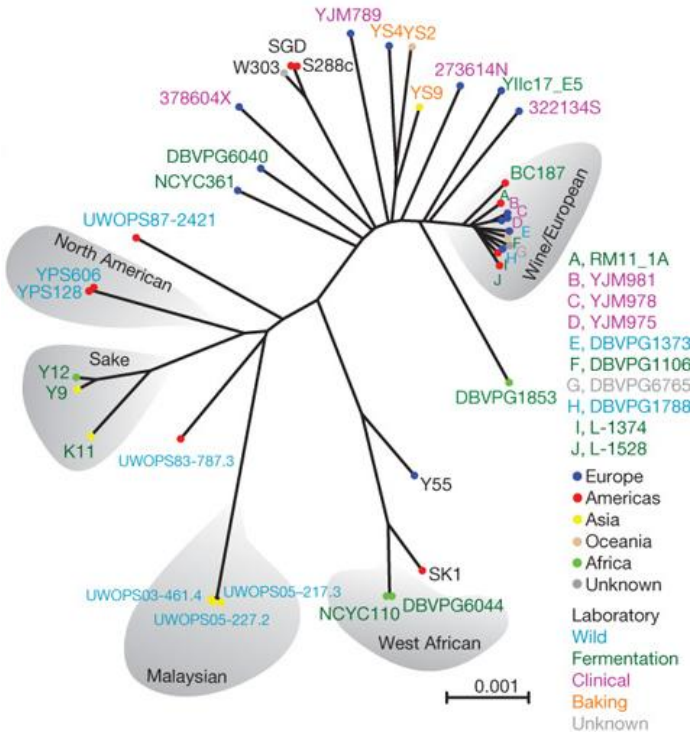

Figure 1: S. cerevisiae tree based on single nucleotide polymorphism, clean lineages are highlighted in gray. Figure adapted from ref [24].

The domestication of yeast

Saccharomyces cerevisiae is one of our most used microorganisms for two reasons. Namely it is used to produce alcoholic beverages and bread. Fermentation creates ethanol from sugar and has long been used for enjoyment, rituals and as a method for preservation [16]. This domestication of *S. cerevisiae* has probably been of vast importance for the spread and adaptation of related strains [24].

The history of wine is of great interest when trying to analyze the phenotypic relation between strains. Growth rate, population size and environmental factors may have had a great influence on the mutation rate within transcription factor genes.
It has been argued that wine making could not possibly have begun before humans started to settle down to agricultural way of life in place of their previous nomadic existence. This is because the grapevine plant (*Vitis vinifera*) bears fruit first two years after they are planted and the grapes can only be picked and crushed during a few days, prior to a long processing time, which would not be compatible with nomadic life. The earliest archaeological evidence of winemaking is from 9000 years ago [16]. The earliest evidence of winemaking in Europe is from a little more than 6000 years ago. This evidence relates to grape juice extraction in Greece during the Classical period [17].
This limits the time period for evolutionary divergence of the European wine yeast strains to a little more than 6000 years. An investigation has shown that *S. cerevisiae* are readily isolated from natural surfaces including wine grapes and wine yard soil. *S. cerevisiae* is much more common in damaged grapes. About one in four damaged fruits carried *S. cerevisiae*, those who carried yeast had as much as 100,000 to 1,000,000 cells [18]. The generation time is harder to calculate, in the lab S288c can have a generation time of 2920 generation/year [18], in the wild growth rates are probably much slower. But domestication (fermentation of wine) probably selects for a fast growth rate and it has been shown that R11-1A has a faster evolutionary rate in laboratory conditions than S288c [19].

Method

Genetic analysis

Sequence data were obtained from Sanger institute's published shotgun sequencing of *Saccharomyces cerevisiae* strains [20]. I choose all the 10 European/wine -group strains for this project and the laboratory strain S288c as a reference.

Table A1: List of used yeast strains, there source and sample location.

Name	Source	Location
DBVPG1373	soil	Netherlands
DBVPG1106	grapes	Austrailia
DBVPG6765	unknown	Unknown
DBVPG1788	soil	Turka, Findland
RM11-1A	vineyard	California
L-1528	fermentation must	Cauqenes, Chile
L-1374	fermentation must	Cauqenes, Chile
YJM975	clinical Isolate	Italy
YJM978	clinical Isolate	Italy
YJM981	clinical Isolate	Italy
S288c	rotting fig	USA

The transcription factors were chosen based on a score for most positive selection of internal IDR's. The scores were taken from a global analysis including 3746 protein coding genes in 37 *S cerevisiae*-strains and 27 *S paradoxus*-strains (J. Nilsson et al, in prep.). The strain sequences were reassembled by inserting single nucleotide polymorphisms (SNP) into the template reference strain of S288c.

I created Neighbour-joining trees with the program MEGA4 [21] using synonymous (data not shown) and non-synonymous mutations in the twelve TFs. IDR locations were noted in the dataset, based on estimates taken from the earlier study by Johan Nilsson.

Deletions and insertions were dealt with by pairwise deletion. I used a bootstrap [22] re-sampling of 5000. Variation in each TF gene was examined using a Z-test to test for neutral evolution (H_A: dN = dS), positive selection (H_A: dN > dS) and purifying selection (H_A: dN < dS).

I analyzed the combined sequence for all the twelve TF by joining them together end to end and creating Neighbour-joining [23] trees for non-synonymous and synonymous codon polymorphisms. This was done in the same way as for the single TF's but with a higher bootstrap level of bootstrap re-sampling of 40.000.

Phenotypic analysis

Cluster analysis of strain functions was performed using data interpreted from growth quantification data published previously [24], the information can be found in its entirety in the supplementary data associated with the paper. Because I could not get the original table data, I interpreted the table by sampling that numeric code for the additive color model BGR. The data was then given a numeric value in order to mirror the scale for growth in the original table. Because of the re-translation of the data there was with all probability a percentage of round off errors and the results are best studied with this in mind. I created a phylogenetic tree based on the phenotypes with the program R, version: 2.9.0 and the tool package R-commander. The tree was created using K-mean cluster analysis using ward's method and Euclidean distance markers. The data from 5 phenotypic tests were not included in the analysis because of missing data. The omitted data were, in the rate column: Serine 1% and Arsenite 5mM, in the efficiency column: Arsenite 5mM and in the column for adaptation: Serine 1% and Arsenite 5mM.

Results

No correlation between gene length and positive selection

First it is proper to exclude other possible factors for the differences observed in the chosen TFs other than evolutionary adaptation. The statistical data of the 12 chosen TFs had few observed common traits. There was a high variance in both sequence length and mutation rate. The shortest TF was YGL254W (900bp) and the longest was YLR278C (4029bp). The frequency of single nucleotide polymorphisms (SNP) could not be explained by sequence length and did only weakly correlate to positive and purifying selection (Table 2). I found no correlation between mutation rate and the ratio between synonymous and non-synonymous mutations (Table 2). Gene length seemed to have a weak positive correlation (correlation coefficient = 0.3) with the non-synonymous vs. synonymous mutation rate (dN/dS) (Fig3), while SNP-amount seems to have the opposite correlation to the dN/dS-ratio. The significance of this is not clear, the effect is very poor and without statistic certainty thus probably negotiable to the project.

Table 2: Statistical data of the 12 investigated TF

TF	No of SNP	No of Coding SNP	No of Non-Coding SNP	bp	Coding mutations/ bp	NonCoding mutations /bp	Coding/ NonConding
YGL254W	5	1	4	900	0.00111	0.00444	0,3
YLR278C	25	6	19	4029	0.00149	0.00472	0,3
YML076C	14	4	10	2835	0.00141	0.00353	0,4
YPL230W	9	3	6	1179	0.00254	0.00509	0,5
YER184C	25	5	20	2385	0.00210	0.00839	0,3
YMR280C	15	7	8	4302	0.00163	0.00186	0,9
YJL056C	25	13	12	2643	0.00492	0.00454	1,1
YMR213W	2	1	1	1773	0.00056	0.00056	1,0
YFL044C	6	1	5	906	0.00110	0.00552	0,2
YPR104C	4	2	2	2811	0.00071	0.00071	1,0
YBR240C	5	1	4	1353	0.00074	0.00296	0,3
YMR037C	4	2	2	2115	0.00095	0.00095	1,0
sum	139	46	93	27231	0.01926	0.04326	7,1

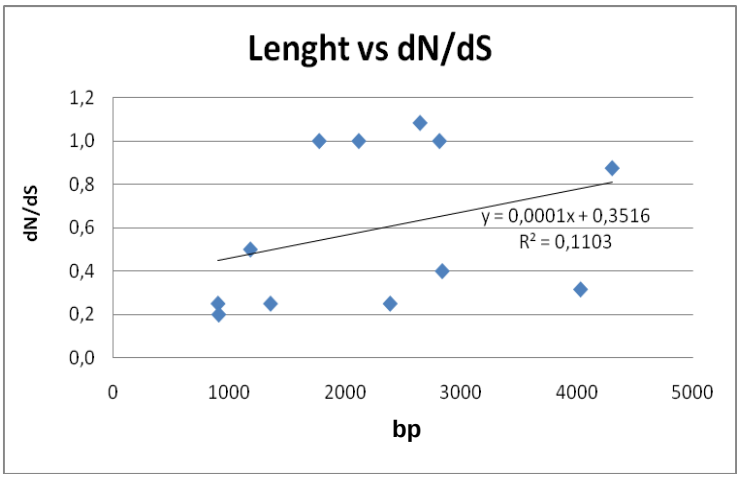

Figure 2: General plot between gene lengths and the ratio between synonymous (dS) and non-synonymous (dN) nucleotide variation

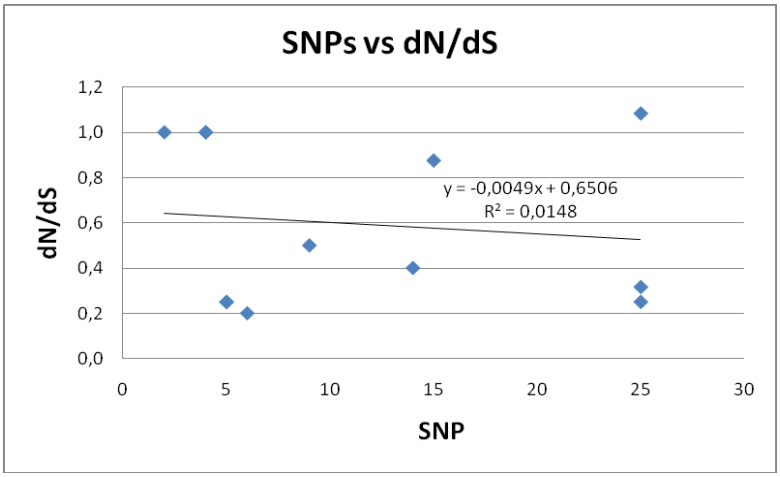

Figure 3: General plot between the amounts of nucleotide variation and its distribution between synonymous (dS) and non-synonymous (dN) nucleotides.

Finding positive selection between the TFs of European and wine yeasts.

I used a neighbour-joining method with only the non-synonymous coding mutations (mutations that causes a change of amino acid in the RNA to protein translation) to create an evolutionary history for each transcription factor. The 12 trees gave a mixed result. The TFs YGL254W, YMR213W and YFL044c had so few coding mutations that the different strains could not be separated from each other. Each had a relative short protein sequence (299aa, 590aa and 301aa, respectively) and only one single non-synonymous mutation separated each of them from the reference strain S288c. YPL230W, YPR104C, YBR240C and YMR037C had only one separating group in the tree, while YML076c (WAR1) had two and YLR27c, YJL056C, YER184C and YMR280C had enough non-synonymous mutations to create multiple distinguishable groups. The observed patterns of single nucleotide variation are represented by the three TFs; YGL254W (Fig4a), YPL230W (Fig4b) and YER184C (Fig4c).

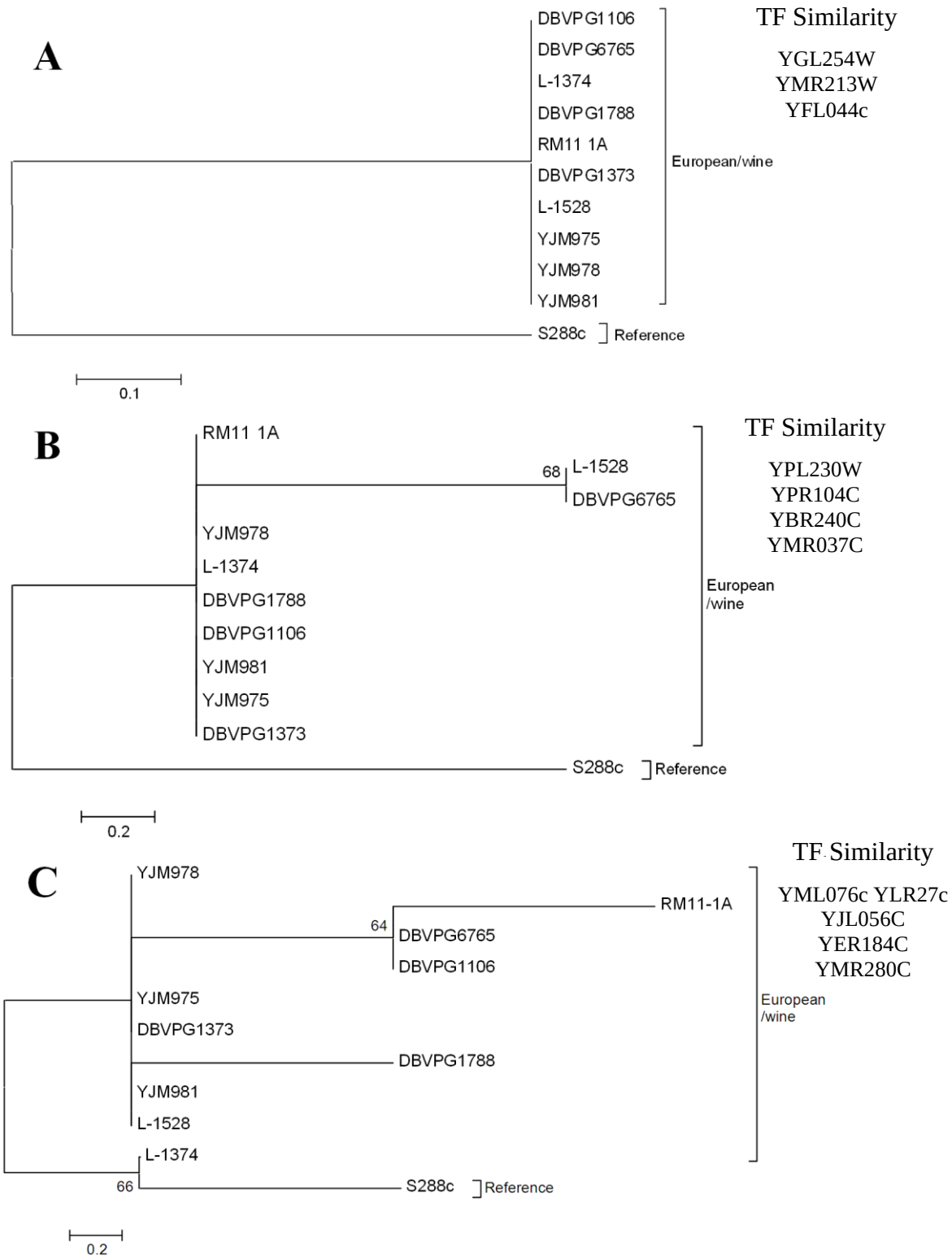

Figure 4 Phylogenetic trees based on the amino acid sequence of the TF's (A): YGL254W, (B): YPL230W, (C): YER184C. They were done by Neighbour-Joining method with a bootstrap of 5000 replicates. Any nucleotide gap dealt with by pairwise sequence comparison, giving a final data length of: YGL254W: 299 positions, YPL230W: 391 positions and YER184C: 793 positions.

Most TFs included too few variable amino acids to create a statistically significant phylogenetic tree. In a codon based Z-test, performed using the Nei-Gojoboris method on the amino acid sequences, there was not statistical support for separation of most of the TF sequences in wine yeasts from the reference strain. There were only 2 TFs that showed significant evidence of positive selection within the European/wine - group, namely YLR278c (Fig5) and YJL056c (Fig6).

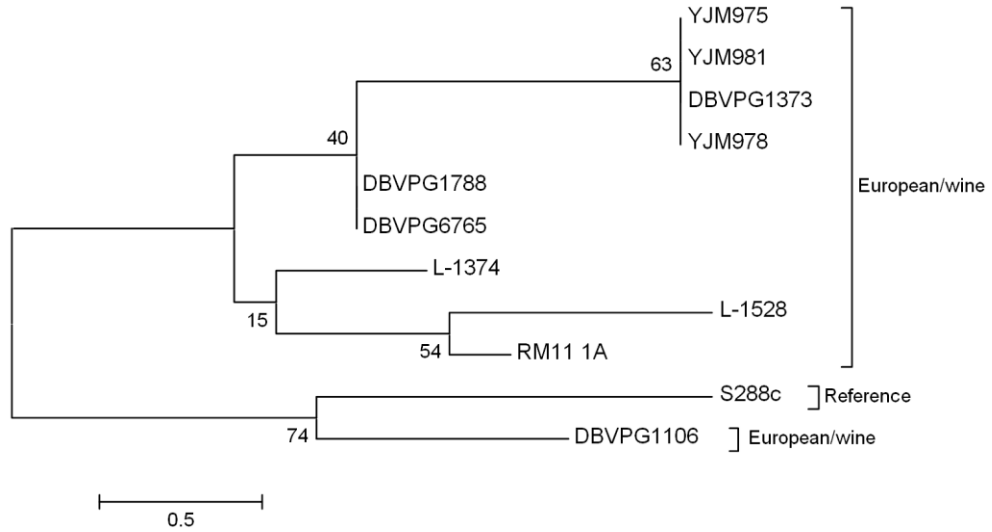

Figure 5: Phylogenetic history based on the TF: YLR278C created with the Neighbor-Joining method by the model of evaluating the differences in amino acid sequence. The tree was tested with a bootstrap of 5000 replicates. The tree is drawn to scale with one amino acid long gap dealt with by pairwise sequence comparison. Final amino acid sequence length: 1342 positions.

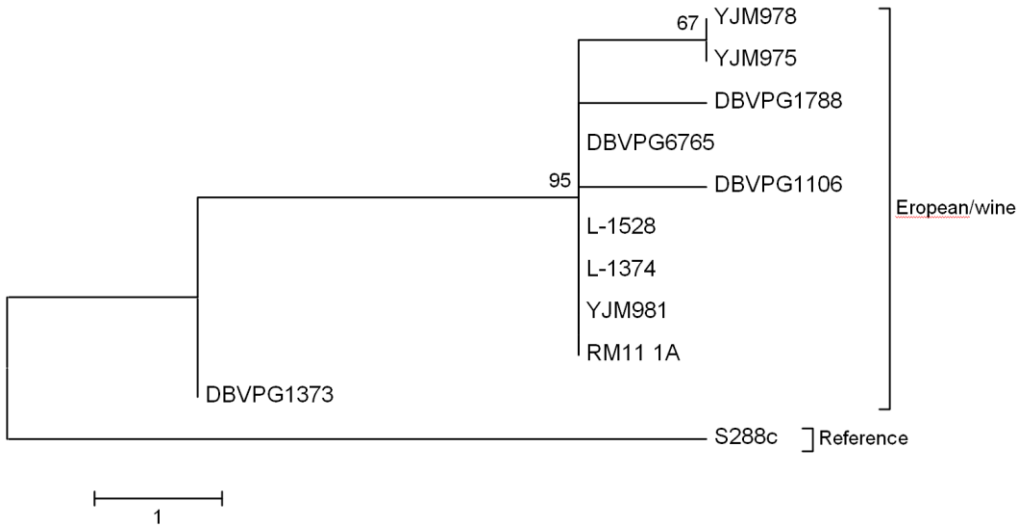

Figure 6: Phylogenetic history based on the TF: YJL056C created with Neighbor-Joining method by the model of selecting for number of differences in the amino acid sequence. The variance was tested with a bootstrap of 5000 replicates. The tree is drawn to a final amino acid sequence length of 880 positions.

YLR278c

Several strains contain YLR278c variants that have a p value under 0.05 in the neutral evolution test (dN = dS). This is seen between the strain DBVPG1106 and a cluster of 4 other strains: DBVPG1373, YJM975, YJM978 and YJM981 (Table 3, Fig7). The same test but with only IDR-sequences gave the same result (data not shown). The less strict positive selection test (dN > dS) gave statistical support for divergence of two more strains, L-1528 and DBVPG1788 (Table 4). Purifying selection was detected for the S288c in relation to all the European/wine strains.

Table 3: Z-test for neutral evolution (dN = dS) in the TF YLR278C, p-values is presented and boxes labelled * are below 5% probability for neutral evolution. Nei-Gojobori (p-distance) method was used with a bootstrap of 1000. Insertions/deletions were dealt with by pairwise deletion.

	S288c	DBVPG1106	DBVPG1373	YJM975	YJM978	YJM981
S288c						
DBVPG1106	**0.00006***					
DBVPG1373	**0.00013***	**0.03705***				
YJM975	**0.00013***	**0.03705***	**1.00000**			
YJM978	**0.00013***	**0.03705***	**1.00000**	**1.00000**		
YJM981	**0.00013***	**0.03705***	**1.00000**	**1.00000**	**1.00000**	

* Significant p-value

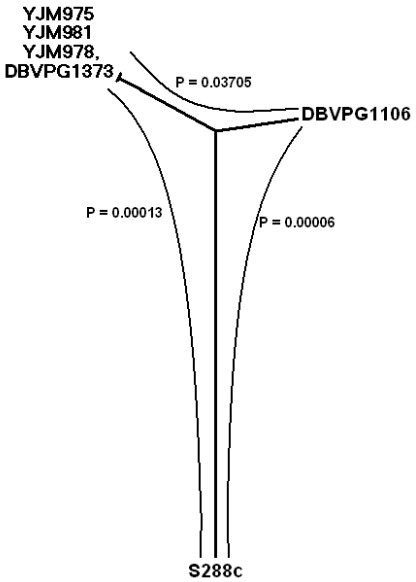

Figure 7: Illustrating significant selection from neutral Z-test in table 3

Table 4: Z-test for positive selection (dN > dS) in the TF YLR278C, p-values is presented and *-marked boxes are above 95% probability of being under positive selection. Nei-Gojobori (p-distance) method was used with a bootstrap of 1000. Insertions/deletions were dealt with by pairwise deletion. Significant positive selection is marked with a * -sign.

	S288c	DBVPG1106	DBVPG1373	DBVPG1788	L-1528	YJM975	YJM978
S288c							
DBVPG1106	1						
DBVPG1373	1	0.01793*					
DBVPG1788	1	0.04409*	0.15993				
L-1528	1	0.04640*	0.04499*	0.08710*			
YJM975	1	0.03705*	1	0.15993	0.04499*		
YJM978	1	0.03705*	1	0.15993	0.04499*	1	
YJM981	1	0.03705*	1	0.15993	0.04499*	1	1

* Significant p-value

YJL056c

For the TF, YJL056C (ZAP1), the positive selection test showed significant support for positive selection changes separating DBVPG1373 from L-1528, R11_1A, YJM978 and YJM981 (Table 5). A restricted Z-test with only IDR sequences showed slightly increased significance of the positive selection test (Table 6). The purifying selection test gave only significant conservation between to the reference strain S288c and the European/wine -group. Also the purifying selection test showed decreased conserved selection against the reference strain.

Table 5: Z-test for positive selection (dN > dS) in the TF YJL056C, p-values is presented and *-marked boxes are above 95% probability of positive selection. Nei-Gojobori (p-distance) method was used with a bootstrap of 1000.

	S288c	DBVPG1373	L-1528	RM11_1A	YJM978
S288c					
DBVPG1373	1				
L-1528	1	0.04292*			
RM11_1A	1	0.04292*	1		
YJM978	1	0.02625*	1	1	
YJM981	1	0.04292*	1	1	1

* Significant p-value

Table 6: Z-test for positive selection (dN > dS) in the TF YJL056C with only IDR-domains, p-values is presented and *-marked boxes are above 95% probability of positive selection. Nei-Gojobori (p-distance) method was used with a bootstrap of 1000.

	S288c	DBVPG1373	L-1528	RM11_1A	YJM978
S288c					
DBVPG1373	1				
L-1528	1	0.03842*			
RM11_1A	1	0.03842*	1		
YJM978	1	0.01894*	1	1	
YJM981	1	0.03842*	1	1	1

* Significant p-value

Several more TFs may contain positively selected mutations but because of the small amount of data, statistic support could not be generated.

36 of 58 combinations in a pair wise Z-test shows evidence of purifying selection

A Z-test were performed estimating purifying selection between all TFs against each other in a pair wise fashion. More than half of the possible pair wise comparisons (36 of 58) showed statistically supported evidence of purifying selection (Table 7). A test using only IDRs gave no significant difference to this result, if anything, the IDR's were a slightly more conserved with 38 of 58 combinations showing significant support for purifying selection. A p-value calculation (the statistical probability that dn<ds by chance) regarding IDR and Non-IDR sequences of all European/wine strains showed statistically probable conservation for almost all strains (Table 8, all data are not shown). There were only two groups that could not in themselves prove positive selection, the collected non-IDR sequences of all European strains, and the non-IDR sequence of ZAP1 (Table 8).

Table 7: Z-test for purifying selection (dN < dS) in all 12 TFs combined, p-values is presented and *-marked boxes are above 95% probability. Nei-Gojobori (p-distance) method was used with a bootstrap of 1000. Insertions/deletions were dealt with by complete deletion.

	S288c	DBVPG-1106	DBVPG-1373	DBVPG-1788	DBVPG-6765	L-1373	L-1528	RM11_1A	YJM975	YJM978	YJM981
S288c											
DBVPG1106	**0***										
DBVPG1373	**0***	**0.24274**									
DBVPG1788	**0***	**0.20361**	**0.07868**								
DBVPG6765	**0***	**0.04817***	**0.06914**	**0.02153***							
L-1373	**0***	**0.02409***	**0.11106**	**0.00551***	**0.06272**						
L-1528	**0***	**0.02118***	**0.12930**	**0.0463***	**0.01754***	**0.08170**					
RM11_1A	**0***	**0.04413***	**0.05960**	**0.03119***	**0.00376***	**0.00656***	**0.01905***				
YJM975	**0***	**0.02499***	**0.06253**	**0.01598***	**0.02257***	**0.20833**	**0.09884**	**0.01709***			
YJM978	**0***	**0.01701***	**0.04808***	**0.02108***	**0.02793***	**0.12969**	**0.15751**	**0.00862***	**0.03578***		
YJM981	**0***	**0.01555***	**0.05761**	**0.0118***	**0.04066***	**0.18366**	**0.23158**	**0.01034***	**0.19036**	**0.31512**	

* Significant p-value

Table 8: Test of purifying selection, calculating rate of coding mutations (dn) and rate of noncoding (ds) mutations between European/wine strains. The significant conserved values are marked with a * -sign. The dn/ds rates has been calculated using Nei-Gojobori method with Jukes-Cantor model in Mega 4.0.2. A bootstrap of 1.000 was used. The P-value is calculated as the probability that the neutral mutations(ds) is only different from the coding mutation (dn) by chance.

All strains	dn	Standard error dn	ds	Standard error ds	p-value	dn/ds	Sequence length (bp)
IDR	0.00062	± 0.00012	0.00323	± 0.00042	0.00000*	0.1919505	20649
non-IDR	0.00027	± 0.00012	0.00339	± 0.00071	0.00001*	0.079646	6546
All European strains							
IDR	0.00035	± 0.00009	0.001371	± 0.00031	0.00171*	0.2552881	20649
non-IDR	0.00026	± 0.00012	0.000506	± 0.00041	0.56154	0.513834	6546
YLR278C							
IDR	0.00074	± 0.00032	0.00388	± 0.00101	0.00306*	0.1907216	19638
non-IDR	0.00045	± 0.00043	0.00422	± 0.00182	0.04409*	0.1066351	948
YJL056C (ZAP1)							
IDR	0.00160	± 0.00044	0.00480	± 0.00148	0.03851*	0.3333333	2325
non-IDR	0.00000	± 0.00000	0.00300	± 0.00305	0.32572	0	312
YER184C							
IDR	0.00125	± 0.00066	0.01453	± 0.00437	0.00265*	0.0860289	1146
non-IDR	0.00038	± 0.00026	0.00414	± 0.00163	0.02247*	0.0917874	1233

* Significant p-value

European strains divide under strong positive selection since separation

To maximize the resolution on the evolutionary history the sequences of all TF were combined, creating a fusion of all phylogenetic trees. The first tree is built on the synonymous differences between the strains (Fig8a). The European strains divide under a rather short period. The TF seams to divide according with their origin (Table 1). The non-synonymous phylogenetic tree (Fig8b) has few things in common to the synonymous, the separation of the strains is greater and in proportion the distance to the reference strain shorter. The YJM-strains are still clustered but have a significant difference between YJM975/YJM978 and YJM981. These mutations can be traced back to only three TFs: YPR104C (FHL1), YML076C (WAR1) and YJL056C (ZAP1). DBVPG1373 is strongly separated from all other strains and restriction to IDR sequences gave no notable difference (Fig 8). DBVPG1373 was most diverged in YJL056C with 3 non-synonymous mutations separating it from the other yeast strains.

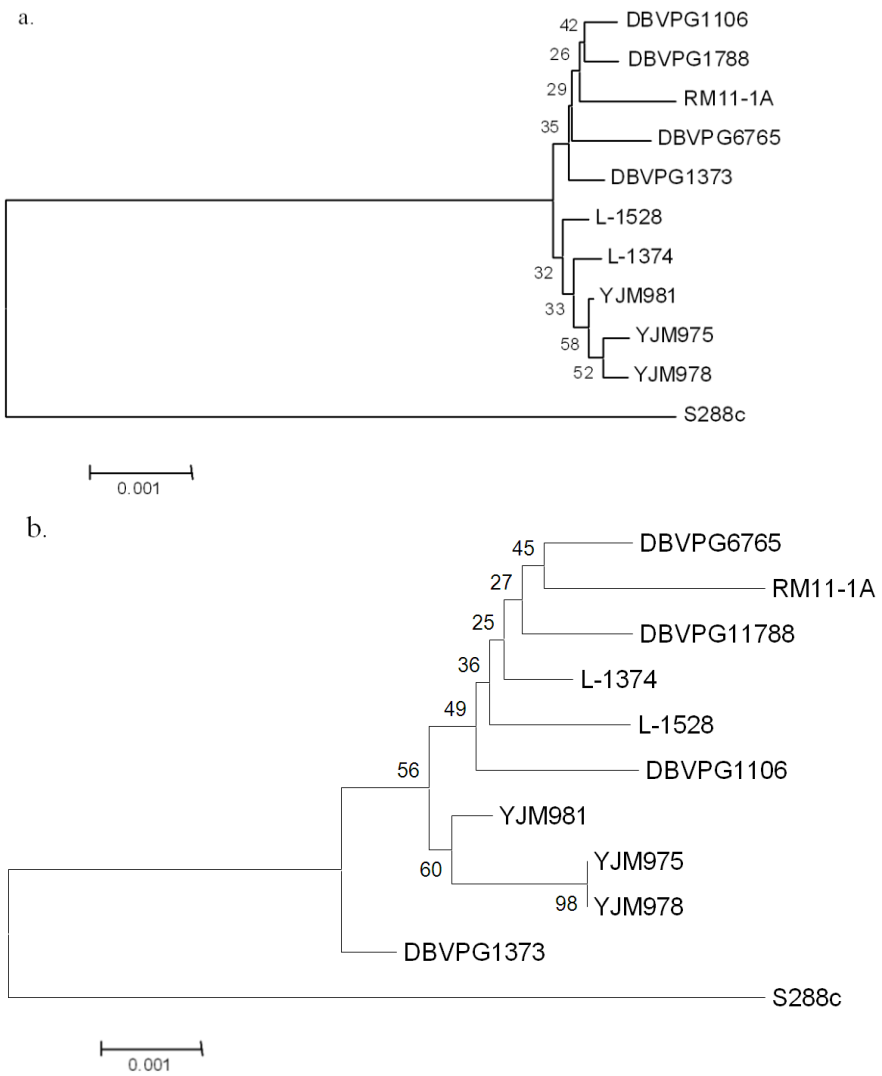

Figure 8: Phylogenetic trees based on the synonymous (a) and non-synonymous (b) differences in 12 TFs. The evolutionary history was deduced using the Neighbor-Joining method with a bootstrap of 40.000 and pairwise sequence comparisons at gaps or deletions. The scale shows distance in p-value and the sequence length was 27195bp.

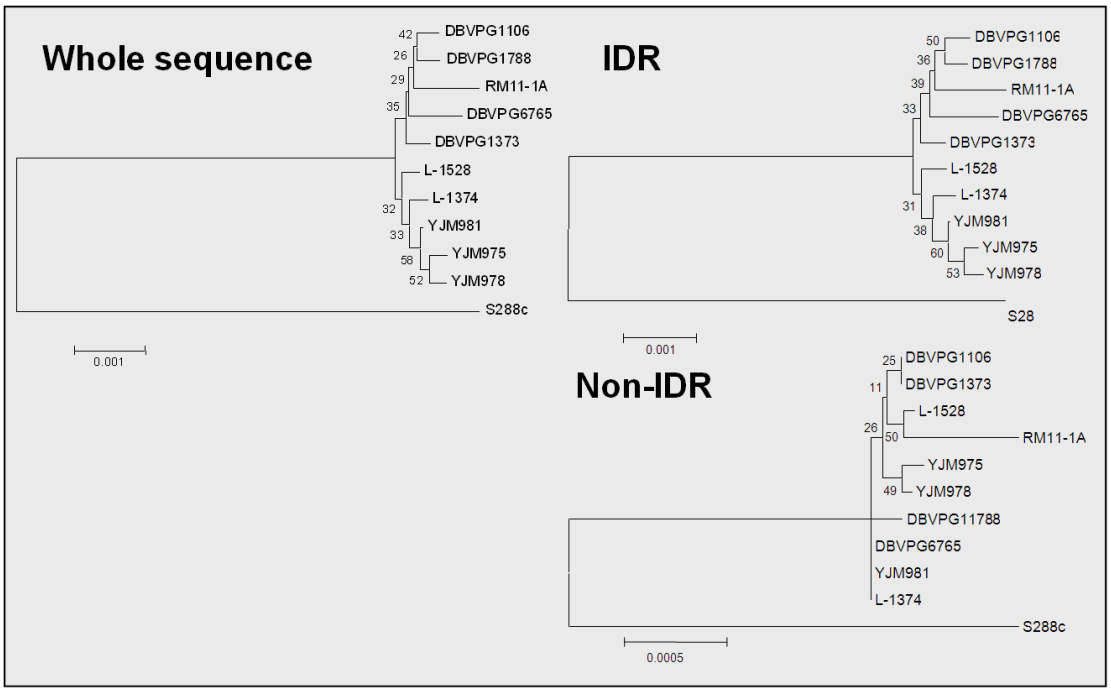

Figure 9: Comparison between all domains versus only IDR and non-IDR domains. Phylogenetic trees based on the synonymous differences in the 12 TFs, including all domains in the first tree and only IDR domains in the second and non-IDR domains in the third. The evolutionary history was deduced using the Neighbor-Joining method with a bootstrap of 40.000 and pairwise sequence comparisons at gaps or deletions. The tree branches are scaled. The total sequence length in the final dataset was 27225 for the whole sequence, 20644bp only including IDR sites and 6581bp of non-IDR-sequence

A Phenotypic comparison divides TFs into two groups

I wanted to track possible phenotypic consequences from the observed mutations. To do this I tried to link single phenotypic traits of the different yeast strains against proposed functions of the TFs. When this gave no conclusive results I did a comprehensive analysis estimating an evolutionary tree based on the growth quantification data, published previously [24]. The results show a separation into two groups, one with DBVPG1788, L-1528, DBVPG1373 and DBVPG1106 and the other with YJM981, YJM978, YJM975, L-1373 and DBVPG 6765 (Fig10). Data for RM11_1A were not included in the growth quantification analysis.

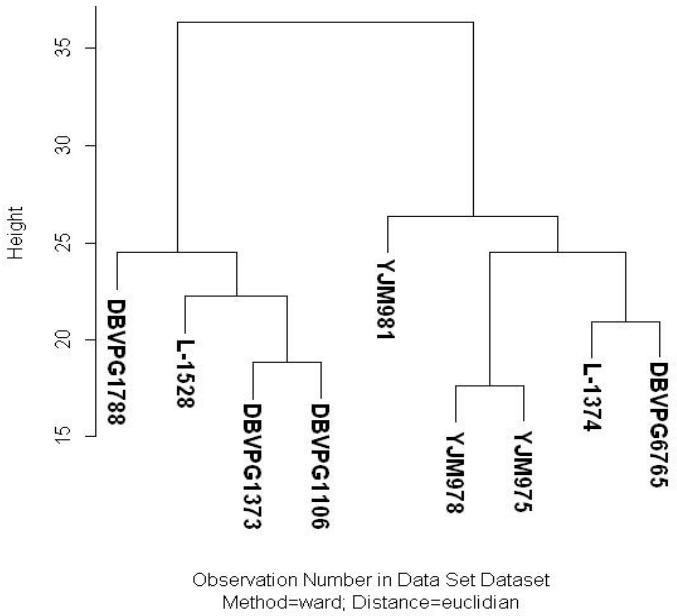

Figure 10: Evolutionary tree based on the phenotypic traits of 9 European yeast strains. The tree was created using K-mean cluster analysis with the method ward and Euclidan distance markers.

Discussion

The Z-test of the combined sequences gave a significant evidence for purifying selection between many of the strains. For example the TF DBVPG1373 showed the least amount of conservation and RM11_1A the most (Table 7). This doesn't mean that TF do not have positive selection, more likely is that there is a lower chance of creating a positive mutation than a negative. All negative mutations make it harder to statistically identify those with positive selection. This is called background selection; when purifying selection in conserved domains can overwhelm adaptive positive selection in closely linked adaptive domains. Also different TFs are probably selected for depending on environmental strain. The TFs I have chosen have an overall importance between *S. cerevisiae* and *S. Paradoxus,* but they may not have as an important part in European/wine yeast evolution. Thus the evolutionary adaptation may go unnoticed.

IDR sequences did not show any higher degree of positive selection (Table 8). Both IDR and non-IDR had an overall purifying selection. But the biggest difficulties were not to get enough IDR sequences but to get enough non-IDR. In the final dataset 20644 out of 27225bp were IDR predicted, that is about 76% of the sequence (Fig9). This combined with a low number of mutations leads to a high degree of statistical uncertainty.

When looking at TFs separately only two of them had significant positive selection. These only showed positive selection between a few selected strains (see strain YLR278c table 4 and YJL056C table 5). The Z-test limited to IDR regions gave no difference in dn/ds-ratio for YLR278c. However the YJL056C showed a slightly decreased p-value, indicating a higher positive selection in IDR domains (Table 6). Other TF showed potential for positive selection. But due to short sequence lengths and few strains the data amount was too small to easily establish statistical significance (Table 2). The sequence length did not seem to correlate to concentration of SNPs nor to positive selection (Fig2, Fig3).

The phylogenetic analyse of the combined TF sequences gave a clue to the evolutionary and genetic history of the European and wine yeast strains. The phylogenetic tree based on the synonymous mutations showed a more recent division between the European strains than that of the Nonsynonymous mutation (Fig8). This indicates that the evolution rate or the positive selection pressure has accelerated after the separation of the European/wine strains. Reasons could be the radical change of the living environments and/or radical changes in population size and/or generation time brought on by domestication.

Two strains were of interest in the non-synonymous tree: DBVPG1373 and YJM981 (Fig8b). DBVPG1373 had a generally high separation from all other strains and showed an especially high rate of differentiating SNPs in the TF YJL056C, a TF directly involved in zinc response.

YJM981 and its closely related strains YJM975 and YJM978 had the best supported difference in the analysis (98%). Two of the mutation separating YJM981 from YJM975 and YJM978 is found in the TF: YPR104C. Because of the otherwise scarce amount of mutation in this TF (no other non-synonymous mutation and 2 synonymous) and it's highly regulating function (rRNA processing) I believe that the two separating mutations in the strains YJM975 and YJM978 might provide the best chance to prove adaptation in TFs.

Wine production has certainly been adapted for yeast, but has yeast adapted for wine? How has modern human societies effected the spreading and adaptation of yeast? Wine fermentation might induce intense competitive cell growth but is this relevant

on an evolutionary level? RM11_1A's extremely high growth rate in the favourable laboratory environment could indicate that [19].

The phenotypic based tree showed little resemblance to either the combined synonymous or nonsynonymous tree (Fig10). Neither did the phenotypes coincide with the clustering of strains in relevant TF. This is of little surprise because of two reasons. A first reason is the huge amount of unconsidered mutations in the rest of the yeast genome. The European/wine strains were chosen because of their close evolutionary relationship but still there are probably hundreds of relevant mutation events separating them apart. A second reason is the crude data created by using an average over all the functions. A more careful approach would be to continue evaluating each phenotypic trait separately. But finding phenotypic consequences for single site mutations is not easy, and to accurately evaluate possible sites it is probably necessary to perform physical experiments.

To further increase our understanding of how TFs affects evolution, and with what tools we facilitate our own evolvement, I believe that deeper research into this field is needed. A more global project, including more TFs and yeast strains, could give us a clearer view of the underlying mechanism of adaptation, and provide a norm for how TFs behave during evolution.

References

[1] Walhout AJ, (2006) "Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping" Genome research Dec;16(12):1445-54.

[2] Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW,
Griswold MD, Chiu W, Garner EC, and Obradovic Z (2001) "Intrinsically disordered protein". Journal of molecular graphics & modelling 19(1):26-59,

[3] Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. (2004) "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life" Journal of Molecular Biology
Volume 337, Issue 3, 26 March 2004, Pages 635-645

[4] Romero P, Obradovic Z, Dunker AK. (2004) Natively disordered proteins: functions and predictions – Applied Bioinformatics ;3(2-3):105-13.

[5] Clay Bracken, Malin M. Young, Keith Dunker. (2001) “Disorder and flexibility in protein structure and function” Pacific Symposium on Biocomputing 6:64-66

[6] http://www.philosophyprofessor.com/philosophers/alfred-north-whitehead.php

[7]. McDonald R.C, Steitz T.A, Engelman D.M. (1979) “Yeast hexokinase in solution exhibits a large conformational change upon binding glucose or glucose 6-phosphate“ Biochemistry Volume 18, Issue 2, 1979, Pages 338-342

[8] Wagner PG, Lynch J V (2008), “The gene regulatory logic of transcription factor evolution” Trends in ecology and evolution, volume 23, issue 7, page 377-385

[9] Conrad M. (1990) “The geometry of evolution” Biosystems 1990, vol 24 pages 61-81.

[10] Eduardo J, Izquierdo J E, Fernando T c (2008) “The Evolution of evolvability in gene transcription networks”, Artificial Life 2008

[11] Huh WK, et al. (2003) “Global analysis of protein localization in budding yeast” Nature 425(6959):686-91

[12] B Akache, K Wu, B Turcotte. (2001) “Phenotypic analysis of genes encoding yeast zinc cluster proteins” Nucleic acids research May 15;29(10): 2181-90

[13] Zhao H and Eide DJ (1997) “Zap1p, a metalloregulatory protein involved in zinc-responsive transcriptional regulation in Saccharomyces cerevisiae” Mol Cell Biol 17(9):5044-52

[14] Kren A, Mammun Y.M, Bauer B.E, Schuller C. Wolfger H, Hatzixanthis K, Mollapour M, Gregori C, Piper P.W, Kuchler K. (2003) “War1p, a novel transcription factor controlling weak acid stress response in yeast. Mol Cell Biol 23(5):1775-85

[15] Schuller C, Mammun Y.M, Krapf G, Schuster M, Bauer B.E, Piper P.W, Kuchler K. (2004) Global phenotypic analysis and transcriptional profiling defines the weak acid stress response regulon in Saccharomyces cerevisiae. Mol Biol Cell 15(2):706-20

[16] Mcgovern, P.E., Zhang, J., Tang, J., Zhang, Z., Hall, G.R., Moreau, R.A., Nunez, A. (2004) Fermented beverages of pre- and protohistoric china. Nature. PNAS. V. 101, NO. 5. P.17593-17598.

[17] Valamoti S.M, Mangafa M, Koukouli-Chrysanthaki Ch, Malamidou D. (2008) “Grape-pressings from northern Greece: the earliest wine in the Aegean?” Antiquity Publications Volume: 81 Number: Pt 311 Page: 54–61

[18] Fay JC, Benavides JA (2005) Evidence for domesticated and wild populations of Saccharomyces cerevisiae PLoS Genet 1(1): eS

[19] Ronald J, Tong H, Brem R (2006) “Evolution rate in laboratory and wild yeast”, Genetics 2006 september, 174:541-544

[20] http://www.sanger.ac.uk/Teams/Team118/sgrp/

[21] Tamura K, Dudley J, Nei M & Kumar S (2007) “MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0” Molecular Biology and Evolution 24:1596-1599.

[22] Felsenstein J (1985) “Confidence limits on phylogenies: An approach using the bootstrap” Evolution 39:783-791.

[23] Saitou N & Nei M (1987) “The neighbor-joining method: A new method for reconstructing phylogenetic trees” Molecular Biology and Evolution 4:406-425.

[24] Liti G, Carter D.M, Moses A.M, Warringer J, Parts L, James S.A, Davey R.P, Ian N. Burt R.A, Koufopanou V, Tsai I.J, Bergman C.B, Bensasson D, O'Kelly M.J.T, Alexander van Oudenaarden, Barton D.B.H, Bailes E, Nguyen A.N, Jones M, Quail M.A, Goodhead I, Sims S, Smith F, Blomberg A, Durbin R, & Louis E.J. (2009) “Population genomics of domestic and wild yeast” nature 458, 337-341.

[25] S Hermann, L Denmat, M Werner, A Santenac, P Thuriaux (1994) “Suppresion of Yeast RNA Polymeras III Mutations by FHL1, a Gene Coding for fork head Protein Involved” Molecular and Cellular Biology, May, 1994 p 2905-2913.